Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation

Sameer Pradhan[1], **Xiaoqiang Luo**[2], **Marta Recasens**[3], **Eduard Hovy**[4] **Vincent Ng**[5], **Michael Strube**[6]
[1]Harvard Medical School, Boston, MA, USA; [2]Google Inc., New York, NY, USA; [3]Google Inc., Mountain View, CA, USA
[4]Carnegie Mellon University, Pittsburgh, PA, USA; [5]University of Texas at Dallas, Dallas, TX, USA; [6]HITS, Heidelberg, Germany

Problem and Objective

- **Problem:** B^3 and CEAF metrics are underspecified for scoring *predicted* (or *system*) mentions as opposed to *key* (or *gold*) mentions, whereas BLANC is only defined for *gold* mentions.
- **Goal:** Clarify the computation of B^3 and CEAF metrics and extension to BLANC for *predicted* (or *system*) mentions (Luo et al., 2014).

Contributions

- Mention manipultation is unnecessary and can produce unintuitive results
- Illustrate the computation of all metrics with a representative example
- Re-scored CoNLL 2011 and 2012 systems
- Make available an open-source, thoroughly-tested reference implementation

Existing Variations

Twinless mentions – Mentions that are either spurious or missing from the predicted mention set (Stoyanov et al., 2009)

- $\mathbf{B}^3_{all}$ (Stoyanov et al., 2009) – All predicted twinless mentions are discarded and recall penalized for those mentions
- $\mathbf{B}^3_{0}$ (Stoyanov et al., 2009) – All predicted twinless mentions are discarded and recall penalized for those mentions
- $\mathbf{B}^3_{ng}$ (Rahmah and Ng, 2009)– All and only those twinless system mentions that are singletons are removed before applying B^3 and CEAF
- $\mathbf{B}^3_{sys}$, $\mathbf{CEAF}^m_{sys}$ (Cai and Strube, 2010) – Twinless key and predicted mentions are manipulated by adding them either from the predicted partition to the key partition or vice versa, depending on whether one is computing precision or recall

Example

$$K = \overbrace{\{a,b,c\}}^{K_1}\ \overbrace{\{d,e,f,g\}}^{K_2}; \qquad R = \overbrace{\{a,b\}}^{R_1}\ \overbrace{\{c,d\}}^{R_2}\ \overbrace{\{f,g,h,i\}}^{R_3}$$

- The key (K) contains two entities with mentions $\{a,b,c\}$ and $\{d,e,f,g\}$ and the response (R) contains three entities with mentions $\{a,b\}$; $\{c,d\}$ and $\{f,g,h,i\}$
- Mention e is missing from the response, and mentions h and i are spurious in the response.

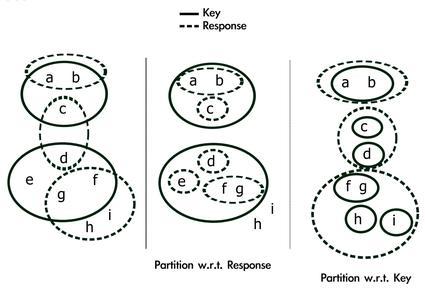

Figure 1 : Partition with respect to key and response

MUC

- Create partitions with respect to key and response as shown in Figure 1

$$R = \frac{\sum_{i=1}^{N_k}(|K_i| - |p(K_i)|)}{\sum_{i=1}^{N_k}(|K_i| - 1)} = \frac{(3-2)+(4-3)}{(3-1)+(4-1)} = 0.40$$

$$P = \frac{\sum_{i=1}^{N_r}(|R_i| - |p'(R_i)|)}{\sum_{i=1}^{N_r}(|R_i| - 1)} = \frac{(2-1)+(2-2)+(4-3)}{(2-1)+(2-1)+(4-1)} = 0.40$$

B^3

- For computing recall each key mention is assigned a credit equal to the ratio of the number of correct mentions in the predicted entity containing the key mention to the size of the key entity to which the mention belongs
- Precision is computed similarly after switching role of key and response

$$R = \frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_r}\frac{|K_i \cap R_j|^2}{|K_i|}}{\sum_{i=1}^{N_k}|K_i|} = \frac{1}{7}\times(\frac{2^2}{3}+\frac{1^2}{3}+\frac{1^2}{4}+\frac{2^2}{4}) = \frac{1}{7}\times\frac{35}{12} \approx 0.42$$

$$P = \frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_r}\frac{|K_i \cap R_j|^2}{|R_j|}}{\sum_{i=1}^{N_r}|R_j|} = \frac{1}{8}\times(\frac{2^2}{2}+\frac{1^2}{2}+\frac{1^2}{2}+\frac{2^2}{4}) = \frac{1}{8}\times\frac{4}{1} = 0.50$$

CEAF

- Get the best scoring alignment between the key and response entities.
- Entity R_1 aligns with K_1 and R_3 aligns with K_2. R_2 remains unaligned.

$CEAF^m$

- $CEAF_m$ recall is the number of aligned mentions divided by the number of key mentions, and precision is the number of aligned mentions divided by the number of response mentions

$$R = \frac{|K_1 \cap R_1| + |K_2 \cap R_3|}{|K_1| + |K_2|} = \frac{(2+2)}{(3+4)} \approx 0.57$$

$$P = \frac{|K_1 \cap R_1| + |K_2 \cap R_3|}{|R_1| + |R_2| + |R_3|} = \frac{(2+2)}{(2+2+4)} = 0.50$$

The $CEAF^m$ F_1 score is 0.53.

$CEAF^e$

- We use the same notation as in Luo (2005): $\phi_4(K_i, R_j)$ to denote the similarity between a key entity K_i and a response entity R_j

$$\phi_4(K_i, R_j) = \frac{2 \times |K_i \cap R_j|}{|K_i| + |R_j|}.$$

$$R = \frac{\phi_4(K_1, R_1) + \phi_4(K_2, R_3)}{N_k} = \frac{\frac{(2\times2)}{(3+2)} + \frac{(2\times2)}{(4+4)}}{2} = 0.65$$

$$P = \frac{\phi_4(K_1, R_1) + \phi_4(K_2, R_3)}{N_r} = \frac{\frac{(2\times2)}{(3+2)} + \frac{(2\times2)}{(4+4)}}{3} \approx 0.43$$

The $CEAF^e$ F_1 score is 0.52.

BLANC

- C_k and C_r – set of coreference links in the key and response respectively
- N_k and N_r – set of non-coreference links in the key and response respectively
- Link between a mention pair m and n is denoted by mn

$C_k = \{ab, ac, bc, de, df, dg, ef, eg, fg\}$
$N_k = \{ad, ae, af, ag, bd, be, bf, bg, cd, ce, cf, cg\}$
$C_r = \{ab, cd, fg, fh, fi, gh, gi, hi\}$
$N_r = \{ac, ad, af, ag, ah, ai, bc, bd, bf, bg, bh, bi, cf, cg, ch, ci, df, dg, dh, di\}$

$$R_c = \frac{|C_k \cap C_r|}{|C_k|} = \frac{2}{9} \approx 0.22$$

$$P_c = \frac{|C_k \cap C_r|}{|C_r|} = \frac{2}{8} = 0.25$$

Coreference score, $F_c \approx 0.23$.

$$R_n = \frac{|N_k \cap N_r|}{|N_k|} = \frac{8}{12} \approx 0.67$$

$$P_n = \frac{|N_k \cap N_r|}{|N_r|} = \frac{8}{20} = 0.40,$$

Non-coreference score, $F_n = 0.50$

BLANC score is $\frac{F_c+F_n}{2} \approx 0.36$.

CoNLL 2011/2012 (English)

System	MD	MUC	B^3	CEAF m	CEAF e	BLANC	CoNLL average
	F_1	F_1^1	F_1^2	F_1	F_1^3		$\frac{F_1^1+F_1^2+F_1^3}{3}$
CoNLL-2011; English							
lee	70.7	59.6	~~68.3~~ 48.9	~~56.4~~ 53.0	~~45.5~~ 46.1	~~73.0~~ 48.8	~~57.9~~ 51.5
sapena	68.4	59.5	~~67.1~~ 46.5	~~53.5~~ 51.3	~~41.3~~ 44.0	~~71.1~~ 44.5	~~56.0~~ 50.0
nugues	69.0	58.6	~~68.8~~ 45.0	~~51.5~~ 48.4	~~39.5~~ 40.0	~~71.1~~ 46.0	~~54.5~~ 47.9
chang	64.9	57.2	~~65.5~~ 46.0	~~54.4~~ 50.7	~~42.0~~ 40.0	~~73.1~~ 45.5	~~56.0~~ 47.7
stoyanov	67.8	58.4	~~61.4~~ 40.1	~~46.1~~ 43.3	~~35.2~~ 36.9	~~60.3~~ 34.6	~~51.9~~ 45.1
CoNLL-2012; English							
fernandes	77.7	70.5	~~71.2~~ 57.6	~~61.9~~ 61.4	~~48.4~~ 53.9	~~77.7~~ 58.8	~~63.4~~ 60.7
martschat	75.2	67.0	~~70.4~~ 54.6	~~59.6~~ 58.8	~~46.6~~ 51.5	~~75.7~~ 55.0	~~61.3~~ 57.7
bjorkelund	75.4	67.6	~~70.3~~ 54.5	~~59.2~~ 58.2	~~45.9~~ 50.2	~~75.8~~ 55.4	~~61.2~~ 57.4
chang	74.3	66.4	~~69.3~~ 53.0	~~58.3~~ 57.1	~~44.8~~ 48.9	~~75.0~~ 53.9	~~60.2~~ 56.1
chen	73.8	63.7	~~68.9~~ 51.8	~~58.1~~ 55.8	~~46.4~~ 48.1	~~74.5~~ 52.9	~~59.7~~ 54.5

Table 1 : CoNLL 2011 and 2012 **official, closed** track performance for English language by top five systems using all predicted information

Conclusion

- Clarified $\mathbf{B}^3$ and **CEAF** metrics and introduced a **BLANC** extension for computing coreference scores given *predicted* (or *system*) mentions.
- Re-scored CoNLL 2011 and 2012 systems
- Released a reference implementation at: `http://code.google.com/p/reference-coreference-scorers`

sameer.pradhan@childrens.harvard.edu